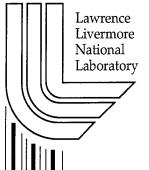
Stochastic Engine: Direct Incorporation of Measurements into Predictive Simulations

R.L. Newmark, R.D. Aines, J.J. Nitao, W.G. Hanley, S.Carle, A.L. Ramirez, S. Sengupta, D. B. Harris

This article was submitted to International Ground Water Symposium, Berkeley, CA, March 25 – 28, 2002

U.S. Department of Energy



August 2, 2001

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This is a preprint of a paper intended for publication in a journal or proceedings. Since changes may be made before publication, this preprint is made available with the understanding that it will not be cited or reproduced without the permission of the author.

This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from the Office of Scientific and Technical Information P.O. Box 62, Oak Ridge, TN 37831 Prices available from (423) 576-8401 http://apollo.osti.gov/bridge/

Available to the public from the National Technical Information Service U.S. Department of Commerce 5285 Port Royal Rd., Springfield, VA 22161 http://www.ntis.gov/

OR

Lawrence Livermore National Laboratory
Technical Information Department's Digital Library
http://www.llnl.gov/tid/Library.html

Stochastic Engine: Direct Incorporation of Measurements Into Predictive Simulations

Robin L. Newmark, Roger D. Aines, John J. Nitao, William G. Hanley, Steve Carle, Abelardo L. Ramirez, Sailes Sengupta, David B. Harris

Lawrence Livermore National Laboratory

Livermore, CA. 94550

Abstract

We are creating a new method of combining disparate types of geologic observations and process simulations. Using Bayesian inferencing and an efficient search algorithm, we obtain a consolidated body of knowledge in the form of multiple configurations and parameter values of the system that are consistent with our existing data and process models. In so doing, we effectively estimate the distributions of both individual parameters and system-wide states, and their likelihood of occurrence. This is in contrast with conventional inversion methods, which produce a single deterministic understanding lacking quantitative information about the distribution of uncertainty. We call this combination of probabilistic evaluation and deterministic process simulators the *stochastic engine*.

Our approach allows the investigators to rapidly improve their understanding of system progress, making it particularly valuable for active processes like injection. The Bayesian inferencing is driven by forward process models that predict data values, such as temperature or electrical voltage, for direct comparison to measured field values. We stage the stochastic searches of possible configurations and run the simplest models, such as lithology estimators, at the lower stages. The majority of possible configurations are eliminated from further consideration by the higher stages' more complex models, such as electrical resistance models for geophysical imaging, or flow and transport models for fluid movement. The approach allows for the continuous augmentation of existing data with newly available information to enhance our understanding and reduce the number of high likelihood configurations. This effectively creates a tool capable of dynamically finding models of underground geological systems that are consistent with all available data. The stochastic engine approach will dramatically increase our understanding of large-scale complex systems and the accuracy of predicting their future behavior under natural or man-made conditions.

Background

A fundamental problem in earth sciences is to determine the properties of an object or process that we cannot directly observe. We use inference and models to extrapolate or interpolate our knowledge, but are fundamentally limited by the inability to inquire over the entire spatial or temporal domain of interest. Here we address this problem by developing a method to simultaneously use many types of data to refine our understanding of complex geologic systems. The stochastic engine uses data (i.e., measurements) together with forward models to determine the likely configurations within a system state space consistent with all available observations. We focus on improving one "base" set of data (or representation of the system) from which other parameters of interest can be calculated using process models. We use the lithology (the general physical characteristics of a rock) as the base representation because it is typically correlated with many parameters of interest such as hydraulic conductivity. It provides a ready means to predict the behavior of the system under forcing events such as fluid injection: when we know the lithology more accurately, we can predict the behavior of the system more accurately. This feedback is central to our ability to acquire enough knowledge about complicated systems to make accurate predictions. The stochastic engine is designed to incorporate everything from the geologists' first field observations to the millions of measurements made during a field operation such as an active remediation project, into an integrated and continuously improving understanding of the base system representation. Rather than generating the "best case" deterministic solution, it determines the most likely solutions, which conventional inversion methods cannot provide. By quantifying uncertainty, it also gives a measure of the value of additional information, and therefore can be used to design experiments as well as determine the value of contributing data.

Due to the flexibility of this method, the stochastic engine can be applied to multidimensional parameters. It can be used to further characterize a system, using existing or newly collected static data, or to improve our understanding of the system dynamics through an analysis of data collected during active processes. As such, some data types that are currently difficult to directly interpret due to temporal and spatial characteristics can be effectively utilized. An example would be the concentration of a tracer as monitored at a down-gradient location over time.

Forward Models

Over the last decade, a ramatic improvement has occurred in numerical modeling capabilities in the earth sciences. Finite difference and finite element models describing fluid and electrical flow (e.g., NUFT-THC (Nitao 1998a, b)) and ParFlow (Tompson et al. 1998)) are now capable of accurately simulating the behavior of large non-isothermal volumes of earth containing multiple phases and subject to complicated forcing functions such as injection of steam. Prediction of changes in chemistry and the actual structure of the formation are now possible in conjunction with mass transport, in several-million-node configurations. A fundamental limitation is that we cannot use our models

effectively without a method to directly identify the millions of input parameters required by a heterogeneous subsurface.

If we have such accurate forward modeling capabilities, why can't we invert for the required details of rock or soil type that control the prediction? The inverse problem is typically ill-posed because the measurements are insufficient to estimate the spatial distribution of subsurface properties in a unique and stable way (Silva, et. al., 2001). The problem lies in the extreme nonlinearity of the problem, the combinatorially massive size of the parameter space, and the fact that many of the observations are highly dependent (i.e., redundant), yielding sparse system matrixes. Another major impediment is the fact that the observations are usually only indirectly related to the property of interest through a complicated function. Utilizing such observations as pumping rates or electrical properties to improve a lithologic model requires a complex model to extract lithologic information. At present, only the simplest geologic problems can be addressed directly in this manner.

The stochastic engine provides a method for resolving this dilemma by iteratively utilizing observations made during an ongoing process to improve understanding. We begin by allowing for a range of properties (such as lithology) and their values. We then use complementary data (e.g. geophysical hydrologic or geochemical) to narrow the of these properties during the course of the process, effectively identifying a smaller and smaller collection of likely configurations. This allows the level of understanding to improve incrementally until sufficient information is obtained to predict and control future activities with an acceptable degree of certainty.

Forward models and prediction

A common problem in earth sciences is that the properties of interest cannot be measured easily in the field at all points of interest. For example, we can measure electrical properties very efficiently, while the actual property of interest is the formation permeability or fluid chemistry. In this case, the interpretation of observed resistivity measurements in terms of permeability requires knowledge of other poorly known parameters. The stochastic engine allows us to directly incorporate field measurements without requiring direct interpretation. It takes advantage of the excellent forward models that describe the process behavior in well-characterized systems. Such forward models are capable of predicting observables in the field. Examples include: temperature at a point, electrical potential between electrodes, or chemical concentration at a screen. While point measurements are commonly used in inversion, long pathlength measurements such as chemical concentrations over time are often available, but, in past analyses, have proven difficult to effectively utilize. The stochastic engine eliminates these difficulties because the comparison is made between measurements and predictions of the same signature properties, and the result is a more complete and accurate use of available information.

The example problem (described later) addresses the question of potential contaminant transport between two wells. The base representation is the lithology, and the observations include both electrical resistance tomography (ERT) data acquired in two

boreholes and a pump test performed between the boreholes. Thus, three data types are involved; lithology, ERT measurements and a pump test. Three computational forward models are used in the analysis (Table 1).

Incorporating Observational Data: Bayesian comparison

One of the fundamental theorems of conditional probability is known as Bayes theorem. The result relates the probability of one event given the occurrence of another (e.g.; A given B occurred), to the inverse conditional probability (e.g.; B given A has occurred) as follows:

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}$$

The probability of B given A has occurred, denoted P(B|A), is called the *posterior* probability, while P(B) alone is the *prior* probability of event B. One manifestation of this theorem has B as a hypothesis being tested and A as an observation pertinent to that hypothesis. Hence, Bayes theorem allows one to revise the initial probability of a hypothesis (P(B)) by incorporating observed data A to produce an updated likelihood of the hypothesis given the observed data (P(B|A)) (which is generally more accurate in light of the new data).

For the current approach, Bayes theorem allows us to model our unknown parameters as random quantities with corresponding probability distributions defined on the space of possible parameter values. This representation of the unknown parameters as random variables is a critical distinction between this approach and prior classical efforts in the earth sciences.

The actual connection of a hypothesis to an observation is made via a forward model: for a possible subsurface configuration, the forward model predicts the values that would be observed by actual measurement. Then, these predictions are compared to the collected data, yielding a likelihood of a match relative to a specified probabilistic error model. The posterior distribution is estimated based upon the results of a sequence of these comparisons which effectively determines the probability that any given subsurface configuration could produce the observed data. By staging these comparisons in a series, we can identify probable configurations using computationally efficient models early in the process. Our most intensive models are only used at the top of the staging process, on configurations that are already known to be consistent with data considered at previous stages (Figure 1).

The configurations that pass all the stages form the collection of possible true configurations of the system. This approach provides a method for incorporating data into our models and assessing uncertainty in our estimates that is not readily available through conventional deterministic approaches. It is an extremely powerful methodology for combining previously known information and newly acquired observations into a single estimate of the probability distribution over the states of the system. By generating

likelihoods of actual lithologies, we can readily employ a wide variety of data in the inference process and use the obtained lithologic likelihoods to guide further investigations. By combining configurations into meta-classes (configurations that are so similar as to behave identically in the field), we can readily deduce whether there is more than one configuration that effectively accounts for both observed and modeled system behavior, and which data will be the most useful in resolving the true configuration from the other high likelihood states.

The staged algorithm is well suited to a number of improvements that we anticipate will be crucial in dealing with complex, three-dimensional problems.

- Any number of stages can be used, involving all data available for the system.
- Data can be added to the algorithm sequentially, as it becomes available.
- The algorithm can be stopped when all available information has been processed, and the newly obtained distribution of the possible configurations can then serve as the basis for processing new data in a subsequent staged algorithm.

Example: Flow in heterogeneous media

We begin with a simple 2-dimensional model in which two boreholes penetrate an alluvial sequence. The model is 26 pixels wide and 40 pixels tall. The question posed is a common one in the environmental arena: whether a pathway exists that could carry contaminants observed in one well to aquifers present in another well (Figure 2). The example problem consists of three lithologic types; an impermeable clay, slightly permeable silt, and a permeable sand. The question is whether or not contaminants present in the uppermost aquifer in the leftmost well can flow to the three lower aquifers present in the rightmost well. In the configuration accepted as the "truth", a permeable hole exists between the aquifer in which the contaminant is observed in the leftmost well and the aquifer immediately below it, presenting a flow path for contaminant transport from left to right into the uppermost aquifer of interest in the rightmost well. However, there are no such paths to the two lower aquifers. Given only the lithology present at the two wells, and assuming a simple layered interpretation, one might conclude that there is no possibility of contaminant transport to aquifers present in the second well.

Using the stochastic engine to address the problem, the selected base representation is lithology because it is typically related to many other parameters of interest. In this case, a completely random search involving three lithologic types with a model of 1040 pixels has 3¹⁰⁴⁰ different possible realizations and hence would require an extremely large number of iterations to fully traverse. The possible number of realizations is greatly reduced by using geostatistical considerations to evaluate configurations that satisfy empirical distributions generated from collected field data (given in Table 2). Using TSIM, we generate a suite of configurations that satisfy these distributions. We choose one at random to constitute the "truth", from which simulated field data will be obtained (Figure 3).

Given the lithologic "truth" and its corresponding simulated ERT and pump test data, the stochastic engine simulation is initiated. The process involves generating a sequence of

Markovian dependent lithologic samples from our prior distribution via TSIM. This sequence is modified by folding in the field data and forward model predictions to effectively produce a sequence of samples from the posterior lithologic distribution. Specifically, assuming a random initial state and n iterations after an initial burn-in phase, the process will have produced a sequence of n samples from the posterior distribution and currently occupy a state X_n . To produce the next posterior sample, a candidate next state Y is generated that is consistent with both our prior lithologic distribution and the current state. The two states X_n and Y are compared in terms of their likelihood of having produced the observed field data and the process transitions to the more likely state via a randomized decision. The critical comparison step is accomplished via two distinct stages described in the following paragraphs.

We next incorporate an additional data type: ERT. Each lithologic unit is expected to have a distinct electrical resistivity. In this simple example, we assign the electrical properties to each given lithology type; however, a more sophisticated approach employing a distribution of properties can be applied. For each configuration, we generate an ERT forward problem using OC4, simulating the presence of 20 ERT electrodes vertically spaced 2 m apart in each of the boreholes. We then compare the predicted ERT signals to those generated by the "truth" configuration. If the latest predicted value matches the "observed" data better than the previous match, it is accepted and becomes the next state of the process. If not, that state is rejected, and another is generated. Using a search method that ensures that the entire parameter space has the potential to be sampled based upon a likelihood measure, we can identify a subset of configurations that are both consistent with the lithologic information and satisfy our criteria for matching the "observed" data. The result is, in effect, a stochastic ERT tomograph, in which the ERT data is used to find possible lithology configurations that fit the data within a given tolerance. This in itself is very useful, as by representing the ERT in lithology space, it permits more direct application to other models such as flow and transport, which are controlled by lithology rather than the electrical properties themselves (Figure 4).

In the final stage we use simulated pump test data. For each configuration that has passed the ERT comparison, we generate a forward pressure test using NUFT-THC to simulate pressures measured at three points in one well based on pumping in the first well. Configurations that pass this stage are consistent with the lithology, ERT and pump test results. By comparing the new results to those obtained by conducting a forward solution for the "truth" case, we rapidly converge towards configurations that include a permeable path or "hole" through which fluid can pass to the uppermost aquifer of interest.

We can quantitatively evaluate the improvement of understanding achieved by incorporating each new data type (i.e., moving through the various stages). The probability of connection from the contaminated zone observed in the left most well to each of the three aquifers of interest at the rightmost well is computed for each stage in Figure 5. The conventional analysis assumes layered media; the probability of a permeable pathway is zero due to the presence of a relatively thick unit of only slightly

permeable silt separating the contaminated zone from the uppermost aquifer, and impermeable clay units separating it from either of the two lower aquifers. However, moving through the three stages indicates the certainty with which the data indicate the potential for a pathway to exist. With the lithology alone, there is nearly a 30% chance of a pathway to the uppermost aquifer and small non-zero probabilities for connection to the lower two. Using both the lithology and ERT, the probability of connection to the uppermost aquifer increases to nearly 0.45, while the probability for connection to the lower two aquifers remain smaller. Including the pump test data requires the presence of a pathway to the uppermost aquifer; the other two appear isolated. Clearly, the pump test data provide a crucial constraint on the solution to this problem.

Summary

We have developed a stochastic engine that combines probabilistic evaluation and deterministic prediction using existing forward models to achieve a quantitatively enhanced understanding of a subsurface system. The stochastic engine uses disparate data (i.e., measurements) together with forward models to determine the configurations of system state space consistent with all available observations. Rather than generating the "best case" deterministic solution, the stochastic engine provides the most likely solutions, which conventional inversion methods cannot provide. By quantifying the uncertainty, it also gives a measure of the value of additional information, and therefore can be used to design experiments as well as determine the value of contributing data.

References

- Carle, S.F., E.M. Labolle, G.S. Weissmann, D. Van Brocklin G.E. Fogg, 1998. Conditional simulation of hydrofacies architecture: a transition probability/Markov approach: in Hydrogeologic Models of Sedimentary Aquifers, Concepts in Hydrogeology and Environmental Geology No. 1, G. S. Fraser and J. M. Davis, eds, Society for Sedimentary Geology, Tulsa (SEPM), OK, 147-170.
- Carle, S. F., G. E. Fogg, 1997. Modeling spatial variability with one and multidimensional continuous-lag Markov chains. Mathematical Geology, OCT, V29 N7:891-918.
- Carle, S. F., 1997. Implementation schemes for avoiding artifact discontinuities in simulated annealing. Mathematical Geology, Feb, V29 N2:231-244.
- Carle, S. F., G. E. Fogg, 1996. Transition probability-based indicator geostatistics. Mathematical Geology, May, V28 N4:453-476.
- LaBrecque, D. J., M. Miletto, W. Daily, A. Ramirez, and E. Owen, 1996, The effects of Noise on Occam's Inversion of Resistivity Tomography Data, *Geophysics*, vol. 61, no. 2,.pp. 538-548.

- Nitao, J.J. (1998a), Reference manual for the NUFT Flow and Transport Code, version 2.0 Lawrence Livermore National Laboratory, UCRL-MA-130651.
- Nitao., J.J. (1998b), User's Manual for the USNT Module of the NUFT code, version 2.0, Lawrence Livermore National Laboratory, UCRL-MA-130653.
- Ramirez, A., W. Daily, A. Binley, D. LaBrecque, and D. Roelant, 1996, Detection of Leaks in Underground Storage Tanks Using Electrical Resistance Methods, *Journal of Environmental and Eng. Geophysics.*, vol. 1, no. 3, pp 189-203.
- Silva, J.B.C., W. Medeiros, and V.C.F. Barbosa, 2001, Pitfalls in non-linear inversion, *Pure and Applied Geophysics, vol. 158*, no. 3, pp 945-964.
- Tompson, Andrew F.B., Robert D. Falgout, Steven G. Smith, William J. Bosl, and Steven F. Ashby (1998) Analysis of subsurface contaminant migration and remediation using high performance computing. *Advances in Water Resources*, V. 22 (3) pp. 203-221.

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National laboratory under contract No. W-7405-Eng-48.

Table 1: Computational forward models used in this analysis.

Code Name	Forward Model	Description	Input	Output
TSIM (Carle et al., 1998, Carle and Fogg, 1996, 1997, Carle, 1997)	Lithology	Generates stochastic realizations of lithology	Hard data (known lithologies) or soft data (inferred lithologies) and model of spatial variability	Grid of lithologic values
OC4 (LaBrecque et al., 1996, Ramirez et al., 1996)	Electrical Resistance Tomography	Computes the electrical potential fields using multiple pairs of electrodes with known currents; solves the partial differential equation for charge conservation	Electrical conductivity field	Electrical voltages
NUFT- THC (Nitao, 1998a,b)	Flow and Transport	Solves the time-dependent, coupled nonlinear system of partial differential equations for energy and mass conservation of air, water, and chemicals in multiple fluid and mineral phases; has back-coupling between chemistry and flow due to changes in permeability	Hydrological and chemical properties (permeability, porosity, mineral surface area, bulk thermal conductivity function, solid heat capacity, gas and liquid relative permeability parameters, capillary pressure parameters, chemical species and reaction parameters); fluid equations of state; initial conditions	Concentration, pressure, temperature, fluid phase void fractions, porosity

Table 2: Geostatistical parameters from field observations

Lithologic Unit	Permeability	Thickness (m)	Width (m)	Fraction (%)	Electrical resistivity (ohm-m)
Clay	Impermeable	1.5	3	25	1.0
Silt	Slight	1.0	3	25	3.16
Sand	Permeable	2.5	12	50	10.0

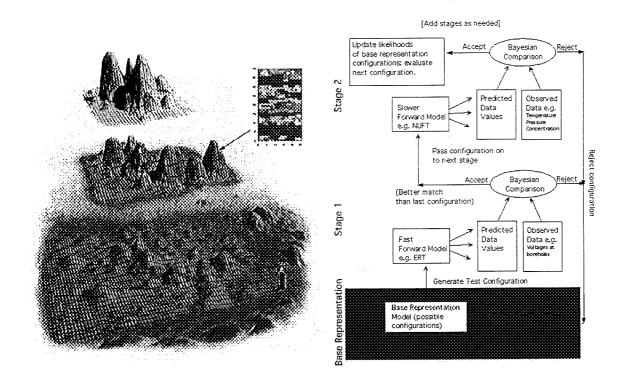


Figure 1. Flow chart for stochastic engine.

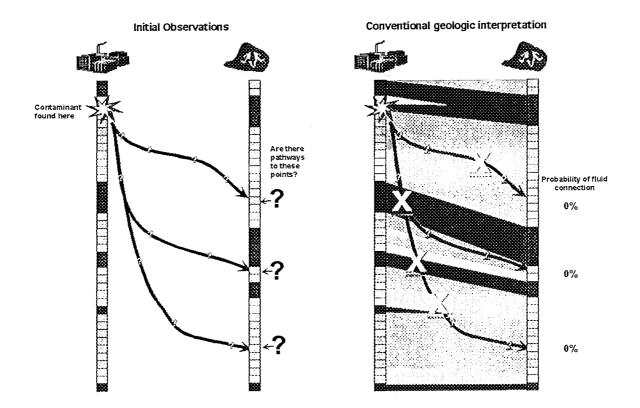
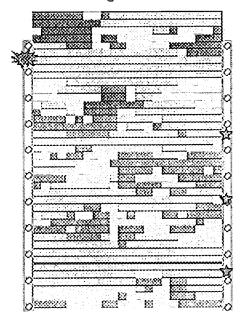


Figure 2. A typical environmental problem is one in which contamination is found in a well: can it spread to nearby aquifers and human receptors? Lithology of nearby wells is similar; are there pathways to permit contaminant transport? Conventional interpretation assigns deterministic layers based on what is observed in the wells. Transport would not be expected to occur at all.

Lithologic "Truth"



Geostatistical Constraints					
Red: impermeable. (Clay)	1.5 m thick 3 m wide 25%				
Yellow: slightly permeable (Silt)	1.0 m thick 3 m wide 25%				
Blue: permeable (Sand - aquifer material)	2.5 m thick 12 m wide 50%				
Geologically, red and yellow tend to be laterally adjacent (75%). Yellow tends to be found above red.					

Simulated Data

Electrical (ERT)

Tressure (NUFT)

Figure 3. Our base representation is lithology: permitted configurations follow field-observed distributions. Following the geostatistical constraints shown above and in Table 2, this synthetic lithologic map generated: it is the "true" configuration. Data used in the stochastic engine only comes from the boreholes and sensors in them. The locations of sensors simulated for the forward predictions are shown.

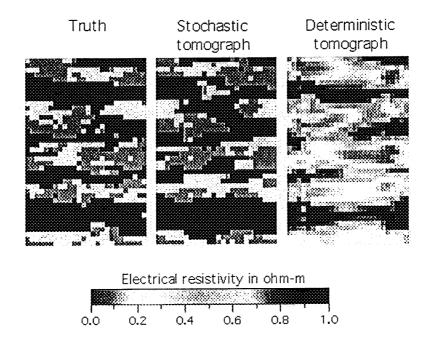


Figure 4. Comparison of ERT tomograph obtained using the standard ERT inversion (deterministic tomograph) with that obtained using the stochastic engine. The lithology chosen to be the "truth" is also shown.

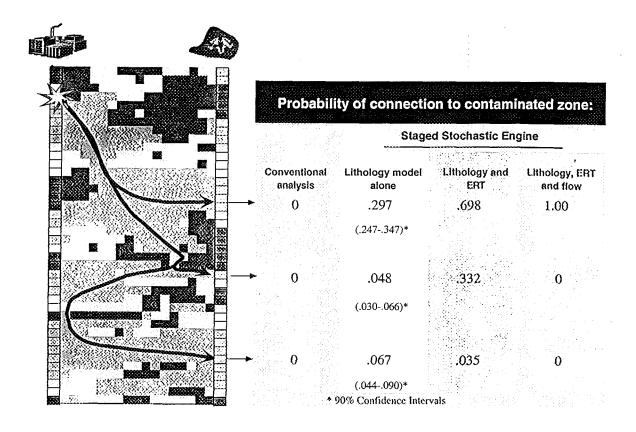


Figure 5. Probability of a permeable connection from the observed contaminant location in one well to aquifers in another well. Note that, by utilizing lithologic, ERT and flow information, we exactly match the "truth" configuration including a connected flow path to the uppermost aquifer, but none to the lower two.

University of California
Lawrence Livermore National Laboratory
Technical Information Department
Livermore, CA 94551